

Relatório de CT-213: Aprendizado por reforço livre de modelo

Henrique F. Feitosa

Instituto Tecnológico de Aeronáutica,
São José dos Campos, São Paulo, Brasil

1 Introdução

Nessa prática, buscou-se implementar os algoritmos de aprendizado por reforço livre de modelo: Sarsa e Q-Learning. Para testar os algoritmos implementados, eles foram submetidos a um processo que os submetia a um MDP que consiste em um tabuleiro unidimensional em que as ações são "STOP", "LEFT" e "RIGHT", sendo que se a partícula realizar o comando "LEFT" na célula da extrema esquerda ela vai para a célula da extrema direita e vice-versa.

Após esse teste, treinou-se uma política para o robô seguidor de linha com os dois algoritmos e analisou-se o resultado.

2 Resultados e Discussão

Os resultados obtidos estão nos testes figuras [1](#) e [2](#).

```

Action-value table:
[[-1.99      -1.      -2.9701   ]
 [-2.96048517 -1.99    -3.92915461]
 [-3.74608059 -2.9701   -4.13688247]
 [-4.49194454 -3.94039893 -5.06243139]
 [-5.13628133 -4.89339226 -4.8937317 ]
 [-4.44187736 -4.62328938 -3.94039877]
 [-3.53809581 -4.42553723 -2.9701   ]
 [-2.96502709 -3.9314936  -1.99     ]
 [-1.99      -2.9701   -1.      ]
 [ 0.        -0.99     -0.99     ]]
Greedy policy learnt:
[L, L, L, L, L, R, R, R, R, S]

```

Figura 1. Mostra o resultado do algoritmo Q-learning

```

[[-9.65364715 -8.57894103 -10.38439963]
 [-10.40009902 -9.46852986 -11.31854977]
 [-11.06463253 -10.39381895 -11.44577623]
 [-11.85643996 -11.32694315 -11.99204891]
 [-12.64278073 -12.34275838 -12.32418257]
 [-11.94021631 -12.05384811 -11.40997038]
 [-11.13154805 -10.96849622 -10.43244015]
 [-10.6346274  -11.33078199  -9.44726591]
 [-9.71319249 -10.34141148  -8.67288193]
 [-7.53201064  -8.65501373  -8.62072075]]
Greedy policy learnt:
[L, L, L, L, R, R, R, R, S]

```

Figura 2. Mostra o resultado do algoritmo Sarsa

Assim, pode-se perceber que o Q-learning tem um pouco mais de exploration em comparação ao Sarsa, que é predominantemente exploitation. Assim, é de se esperar que os resultados achados pelo Q-learning encontrem valores menores. Os resultados da política do robô seguidor de linha estão representados pelas figuras 3,4,5,6,7,8,9 e 10.

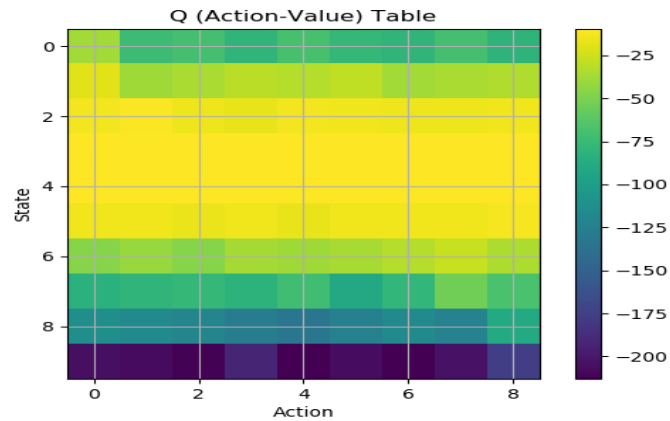


Figura 3. Mostra o os valores para cada estado e cada ação tomada no algoritmo Sarsa

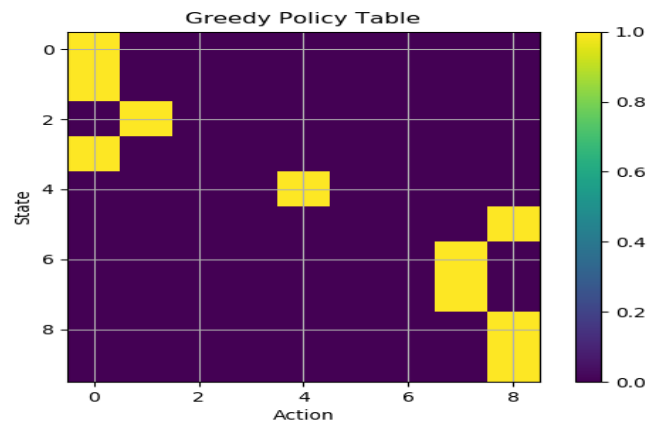


Figura 4. Mostra a política encontrada pelo para cada estado e a ação tomada pelo algoritmo Sarsa

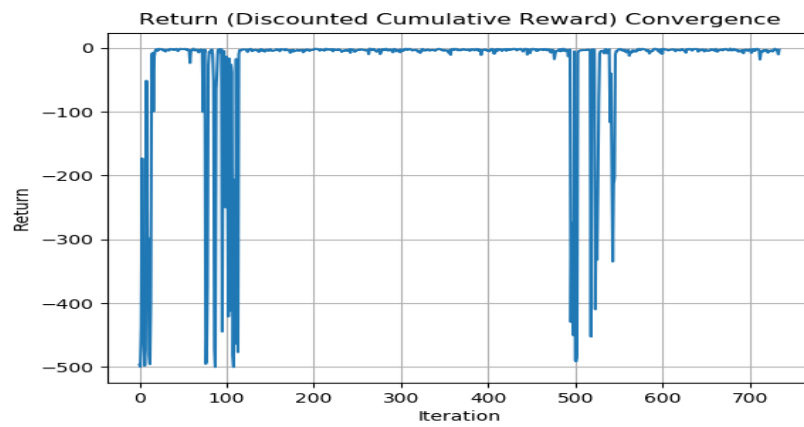


Figura 5. Mostra a convergência encontrada pelo algoritmo Sarsa ao longo das iterações

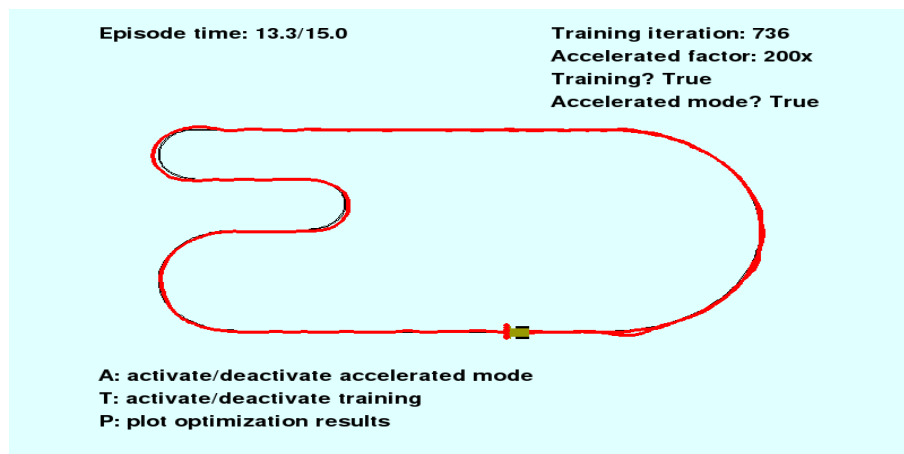


Figura 6. Mostra a política ótima encontrada do algoritmo Sarsa

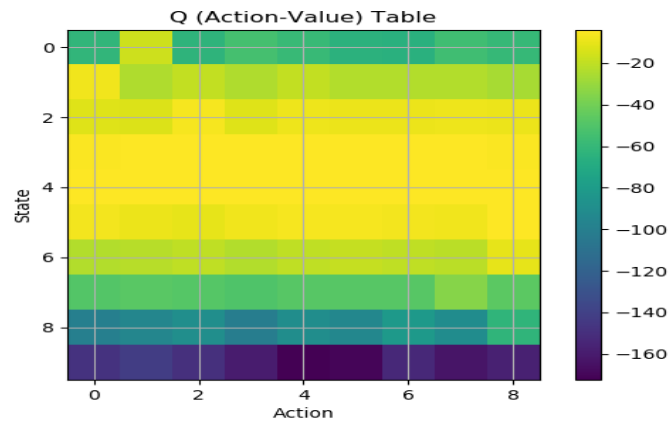


Figura 7. Mostra o os valores para cada estado e cada ação tomada no algoritmo Q-learning

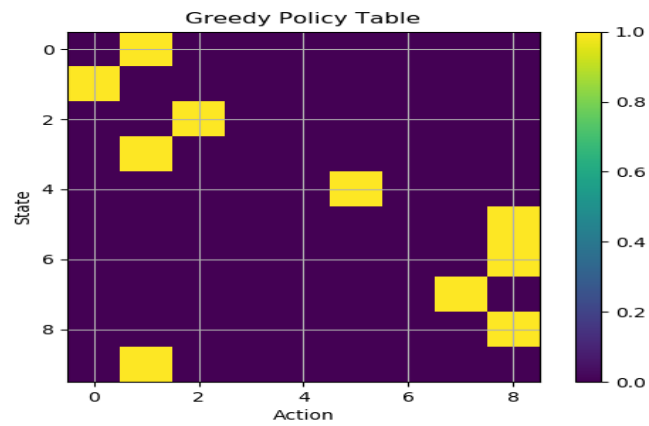


Figura 8. Mostra a política encontrada pelo para cada estado e a ação tomada pelo algoritmo Q-learning

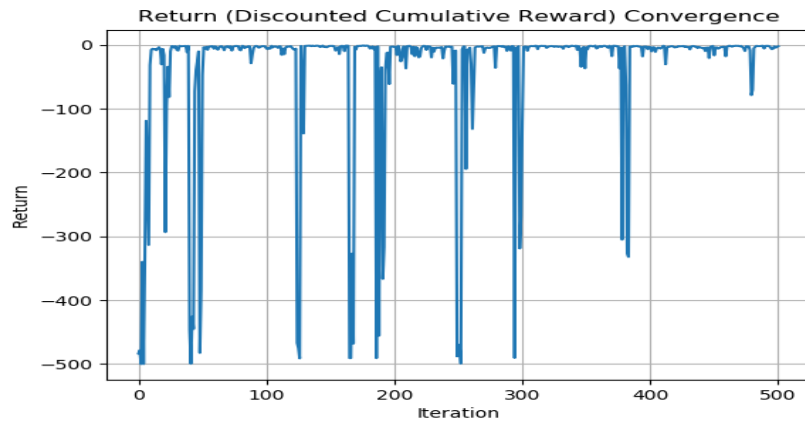


Figura 9. Mostra a convergência encontrada pelo algoritmo Q-learning ao longo das iterações

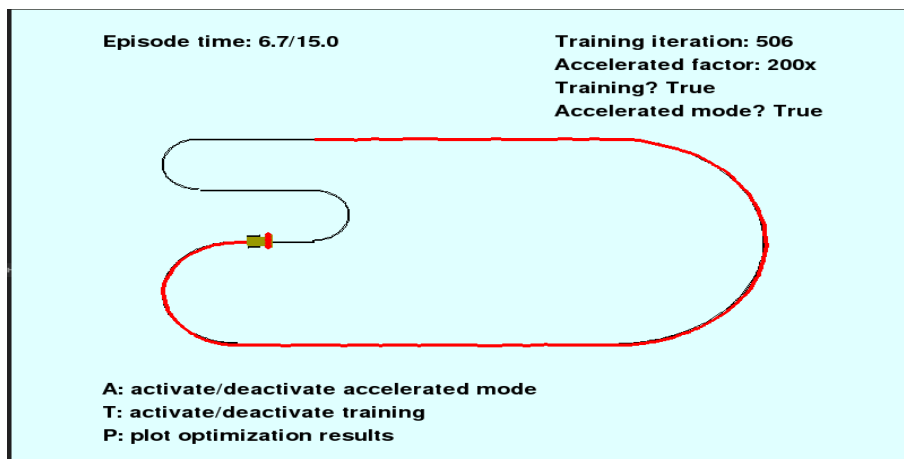


Figura 10. Mostra a política ótima encontrada do algoritmo Q-learning

Pelas figuras 5 e 9, pode-se perceber que a convergência dos parâmetros foi aceitável, pode-se confirmar isso pelas políticas mostradas nas figuras 10 e 6 que mostram o caminho encontrado pelo robô depois de um número considerável de informações. Além disso, é notável perceber que durante as iterações, o algoritmo Q-learning muitas vezes ia por caminhos não tão bons, visto que o robô saía da linha facilmente, isso confirma sua característica forte de exploration.